J.-P. BENZÉCRI
F. BENZÉCRI

Le codage linéaire par morceaux : réalisation et applications

Les cahiers de l'analyse des données, tome 14, nº 2 (1989), p. 203-210
<http://www.numdam.org/item?id=CAD_1989__14_2_203_0>

Les Cahiers de l'Analyse des Données
Vol XIV 1989 n°2 - pp. 203 210.

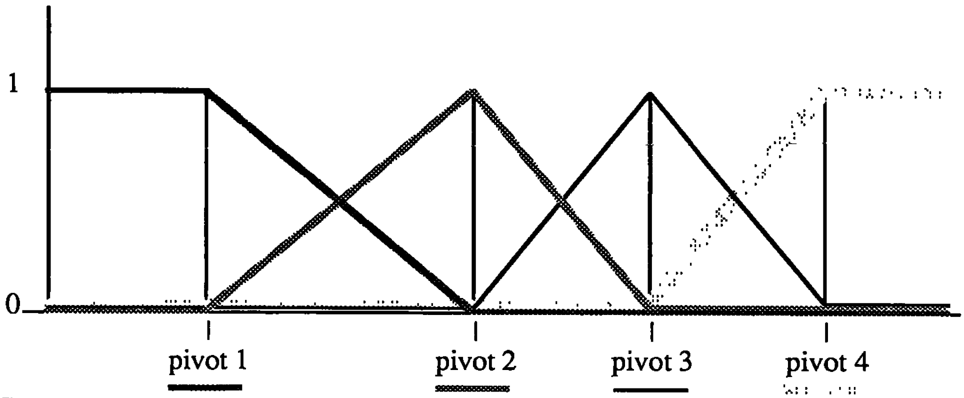
LE CODAGE LINÉAIRE PAR MORCEAUX: RÉALISATION ET APPLICATIONS

[CODAGE LIN.]

J.-P. & F. BENZÉCRI

Après avoir rappelé au §1 le type de codage flou que nous appelons ici "codage linéaire par morceaux", on indique au §2 comment ce codage a été introduit en option dans le programme 'zrang'; puis on montre au §3 quelles applications s'offrent à ce codage ; et on termine sur un exemple de traitement de données.

1 Le codage linéaire par morceaux



Recodage barycentrique d'une variable en 4 modalités autour de 4 valeurs pivot : on a figuré avec des traits différents les valeurs prises par chacune des modalités.

Ainsi qu'il apparaît sur le graphique, il s'agit de recoder une variable dont on suppose que les valeurs v sont des nombres qu'on peut assimiler à une abscisse sur l'axe horizontal. (Dans les applications, il se pourra que v varie continuement, ou prenne seulement un nombre fini de valeurs entières, etc...). Les modalités sont créées autour de valeurs repères appelées ici *pivot*.

Sur la figure, on a choisi d'illustrer le cas où il y a quatre modalités: {mod1, mod2, mod3, mod4}. Si la variable v est inférieure ou égale à pivot 1, mod1 vaut 1; et les trois autres modalités sont nulles; entre pivot 1 et pivot 2, mod1 décroît linéairement de 1 à 0; et elle reste nulle ensuite. La modalité mod2 est nulle jusqu'à pivot 1; elle croît linéairement de 0 à 1 entre pivot 1 et pivot 2; décroît linéairement de 1 à 0 entre pivot 2 et pivot 3; et reste nulle ensuite. De même, la modalité mod3 ne diffère de 0 que si v est comprise entre pivot 1 et pivot 3; le maximum, 1, étant atteint sur le pivot 2. Enfin la dernière modalité mod4, nulle jusqu'à pivot 3 croît linéairement jusqu'à 1 entre pivot 3 et pivot 4; et garde cette valeur ensuite.

On remarquera qu'il n'y a jamais plus de deux modalités différentes de 0; que les valeurs des modalités sont toutes positives ou nulles; et que le total en est 1, quelle que soit la valeur v.

2 Modifications au programmes 'zrang' et 'planF'

Nous indiquons comment réaliser un codage linéaire par morceaux à l'aide de la version actuelle du programme 'zrang'; et signalons au passage les options introduites dans nos programmes depuis la publication des notices [NOT. CRÉ. TAB.] et [NOTE CORR. CAH] (in *CAD*, Vol XIV, n°1).

2.1 L'option barycentrique de 'zrang'

Reportons-nous au §2.2 de la notice [NOT. CRÉ. TAB.].

Si l'on a pris l'option 'D' de 'zrang', s'offrent maintenant à l'utilisateur 3 options au lieu de 2. En effet, s'affiche la question:

'faut il créer un tableau booléen (B), barycentrique (Q)
ou un tableau des numéros des modalités (M) '

La réponse 'Q' correspond à ce que, au §1, on a appelé codage linéaire par morceaux. Voici pourquoi on a choisi le terme de barycentrique. Supposons, pour reprendre la figure du §1, qu'il y a quatre modalités, définies par la donnée de leurs pivots {pivot 1,... , pivot 4}. Les quatre pivots divisent l'axe de variation de v en 5 intervalles consécutifs (les intervalles extrêmes étant, plus précisément, des demi-droites). Sur les intervalles extrêmes, v est simplement recodée, respectivement, suivant {1, 0, 0, 0} et {0, 0, 0, 1}; sur chacun des trois intervalles internes, v tombe entre deux pivots, et peut être considérée comme le barycentre (ou centre de gravité) de ces deux pivots munis de masses appropriées ayant pour somme 1. Ces masses sont précisément les valeurs attribuées aux deux modalités non nulles correspondantes.

Relativement aux options 'M' ou 'B', le dialogue de création des modalités est peu modifié dans l'option 'Q'. Dans toutes les options, on demande d'abord le nombre de modalités choisies. Puis pour chaque modalité, on demande de spécifier un sigle et une valeur (définie par un rang dans la suite des valeurs effectivement prises par *v*). Mais tandis que pour un codage en (0,1) il est inutile de demander une borne pour la dernière modalité qui s'étend, en effet, jusqu'aux plus grandes valeurs de *v*, il est ici nécessaire de fixer un pivot pour chacune des modalités, *modn*, de la première à la dernière. D'autre part la question est modifiée; on demande:

'le RANG de la valeur pivot de *modn* est '

(au lieu de demander: 'le RANG de la valeur maxima de *modn* est ').

Ce qui vient d'être dit concerne le découpage des variables, en dialogue, sans utiliser de bornes préétablies. Si l'on utilise un fichier de bornes 'Dcodx', celui-ci a même structure qu'il s'agisse d'un codage en (0,1) ou d'un codage flou (barycentrique). On verra au §3 qu'on peut réaliser des codages très divers en utilisant l'option 'Q' avec un fichier 'Dcodx' approprié.

2.2 L'option 'ranger' de 'zrang'

Reportons nous au §3.3 de la notice [NOT. CRÉ. TAB.]. On voit que, (pour toute variable *v* dans le cas de l'option 'R', "ranger", et pour les variables pour lesquelles le nombre de modalités est zéro, dans le cas de la sous-option 'M', "numéro de modalité" de 'D', "découper en classes"), est crée une colonne où est inscrit, sur la ligne *i*, le rang qu'occupe la valeur initiale k(i,v) dans l'ensemble des valeurs prises par la variable *v*, triées dans l'ordre croissant.

Or, en toute rigueur, le rang n'est bien défini que si les valeurs prises par *v* pour les différents individus *i* sont toutes distinctes entre elles. Sinon, il y a des blocs consécutifs comprenant éventuellement plusieurs valeurs identiques. Aux individus d'un tel bloc, il semble naturel d'attribuer un même rang qui sera la moyenne des numéros des places occupées par les extrémités bu bloc.

Prenons un exemple:

numéro de i	1	2	3	4	5	6	7	8	9
valeur de v	73	84	80	73	120	58	84	40	84
rang pour v	3.5	7	5	3.5	9	2	7	1	7
double de r	7	14	10	7	18	4	14	2	14

on voit que, si un bloc comprend un nombre impair de valeurs, le rang moyen attribué au bloc n'est pas un nombre entier. Le format entier pouvant intéresser certaines applications ultérieures, on a préféré remplacer désormais, dans tous les cas, la valeur du rang (ou du rang moyen) par la valeur double.

2.3 Représentation des individus sur les graphiques

L'occasion se présente de signaler une option nouvelle du programme 'planF' (cf. [NOTE CORR. CAH] §2.2). À propos de chacun des ensembles à représenter, *i*, *is*, *iq*, *j*, ..., s'offre à l'utilisateur un choix:

'faut-il figurer les élements *ix* par leurs sigles(S)
ou par un caractère unique(U) '

Et, si la réponse est 'U', il reste à fixer le caractère en répondant à:

'le caractere choisi pour *ix* est '

Cette option permet d'abord d'obtenir une image de densité d'un nuage formé de nombreux points, en figurant chacun de ceux-ci par une lettre unique ou un signe de ponctuation. On peut éventuellement apporter ainsi des informations précises, (plus intéressantes, en fait, que les sigles mêmes des individus), si l'on a créé (par le programme 'soustab', cf. [NOTE CRÉ. TAB.] §5) des sous-fichiers comprenant les individus rentrant dans une certaine caractéristique. C'est ainsi que dans [NOTES MOTS] (in *CAD*, Vol XIV, n° 1, p. 92) on a présenté les sous-nuages des électeurs de Chirac et de Lajoinie.

3 Diverses utilisations du codage linéaire par morceaux.

3.1 Codage de variables continues

On peut préférer au codage en (0,1) le codage barycentrique si, les individus étant peu nombreux, on hésite à multiplier les modalités: le codage barycentrique servira comme un compromis entre divers découpages en classes que suggère l'histogramme. Si 2 individus *i* et *i'* occupent une position excentrique du côté du minimum de la variable *v*, en plaçant le pivot 1 entre ces individus et le reste du lot, puis le pivot 2 un peu au-delà, on pourra obtenir que la modalité mod1 vaille 1 pour *i* et *i'*, puis environ 0,5 pour deux ou trois autres individus; ce qui assurera sans doute à mod1 une place moins sensible aux particularités de l'échantillon considéré que si on l'avait réservée aux seuls individus *i* et *i'*. Il importe toutefois de noter ici que, pour placer le pivot 1 entre deux individus, il faudra en marquer la place sur l'axe *v* dans le fichier Dcodx.

3.2 Recodage de variables déjà codées en (0,1)

Il est commun de recoder suivant 3 modalités {mod1, mod2, mod3} une variable déjà codée en (0,1) suivant 5 modalités {m1, m2, m3, m4, m5}; avec le schéma ci-joint, donné pour 5 individus ix rentrant dans chacune des modalités mx.

modalité	m1	m2	m3	m4	m5	mod1	mod2	mod3
i1	1	0	0	0	0	1	0	0
i2	0	1	0	0	0	1/2	1/2	0
i3	0	0	1	0	0	0	1	0
i4	0	0	0	1	0	0	1/2	1/2
i5	0	0	0	0	1	0	0	1

Ce recodage sera réalisé si, après avoir créé un fichier M de numéros de modalités correspondant au découpage de *v* en 5 classes, on applique au fichier

M un recodage linéaire par morceaux avec pour valeurs des pivots 1, 3 et 5. On imaginera aisément comment recoder de même 7 modalités sur 4; ou, en général, 2n - 1 sur n; ou encore 4 modalités sur 3 (en plaçant les pivots en 1, 5/2 et 4).

Il importe de noter qu'il est aisé de ne recoder effectivement que certaines variables. En effet si, pour une variable *v* découpée en *n* classes, on place *n* pivots, ayant pour abscisse {1, 2, ..., n}, le recodage linéaire par morceaux produira, pour cette variable le codage usuel en (0,1) suivant *n* modalités.

3.3 Dédoublement et codage par rangs dédoublés

En fixant seulement deux pivots, situés chacun à l'une des extrémités de l'intervalle de variation d'une variable numérique *v*, on obtient par codage barycentrique deux modalités complémentaires qui ne sont autres que *v* et -*v* recadrées entre 0 et 1 et 1 et 0 respectivement: c'est ce qu'on appelle d'ordinaire un codage par rangs dédoublés.

En appliquant ce codage à un tableau 'R' de rangs (ou à une colonne du tableau 'M' pour laquelle on a demandé zéro modalité, cf. *supra* §2.2) on aboutit à un codage par rangs dédoublés. On obtient un résultat peu différent si on opère sur une colonne du tableau 'M' pour la quelle on a demandé un nombre élevé de modalités; en sorte que le numéro de la modalité constitue une approximation, par degrés entiers, d'un codage linéaire de la variable initiale.

Nous pensons que l'utilisateur a ainsi, dans l'actuelle version de 'zrang', le moyen de réaliser, en un ou deux passages, la plupart des codages usuels, même s'il désire coder les variables suivant des formats différents. Certes, les tableurs se prêtent également à toute sorte de codage; mais, sur un ensemble de quelques centaines d'individus, le temps de calcul par tableur nous paraît long. Quant à écrire un programme spécifique de codage, tous les utilisateurs de l'analyse des données n'y sont pas préparés.

4 Une application du codage barycentrique: analyse des performances de matières thermoplastiques

4.1 Description des données

Sous le titre "Les dix thermoplastiques qui battent des records", P. Personnaz entretient les lecteurs de *Technologie*, (supplément de Janvier 1989 à l'*Usine Nouvelle*), des propriétés de matières nouvelles qui rivalisent avec les alliages métalliques spéciaux. L'article contient quelques données numériques qui nous ont paru mériter d'être analysées, même s'il ne s'agit pas d'une description mécanique complète des nouveaux composés.

Chaque produit est considéré, en principe, sous deux formes: non renforcé et renforcé de 30% de fibres de verre, (ce qui améliore la plupart des qualités). Nous utilisons des sigles en minuscules pour les produits simples; et en

Usine Nouvelle, Technologie, Janvier 1989, superplastiques.

16	ps	PS	pei	PEI	pes	PES	pcl	PCL	hta	HTA	PPS	pai	pek	PEK	pa6	PA6
Tflx	174	180	200	210	200	230	170	230	230	252	270	278	150	315	80	255
Tvll	150	170	170	170	180	200	160	220	200	200	210	220	250	250	100	130
Rtrc	65	108	105	160	90	140	230	170	85	135	160	195	100	160	90	170
Rflx	27	75	33	83	30	85	90	160	25	85	145	46	40	100	25	85
Rchc	70	85	50	100	84	80	270	130	123	59	80	134	80	95	40	120

majuscules pour les produits renforcés. Les produits pour lesquels on dispose des valeurs de 5 variables sont:

ps : polysulphone; **pi** : polyétherimide; **pes** : polyéthersulfone; **pcl** : polymère à cristaux liquides; **hta** : variante améliorée de pes; **pps** : polysulfure de phénylène; **pai** : non décrit; **peek** : polyétheréthercétone; **pa66** : non décrit.

Des 5 variables mesurées, deux sont des températures maxima supportées et les trois autres sont proprement des grandeurs mécaniques:

Tflx : température de déflexion sous charge; caractérise la tenue à des pointes de température de courte durée.

Tvll : température de service en continu; tient compte du vieillissement éventuel du polymère.

Rtrc : résistance en traction, en Méga Pascal;

Rflx : Module de flexion, en Giga Pascal (multiplié par 10 dans notre tableau)

Rchc : Résistance au choc, en Joule/m; essai américain Izod sur échantillon entaillé.

4.2 Codage des données

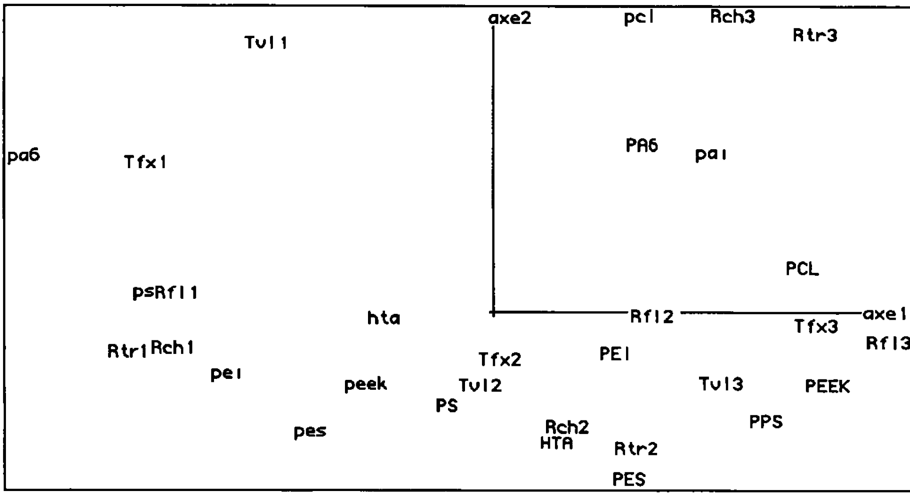
On a d'abord utilisé le programme 'zrang' pour transposer le tableau qu'il nous avait paru plus commode de saisir variable par variable; puis le même programme 'zrang' a servi pour découper une première fois les variables en mode 'Q' (barycentrique) avec des bornes fixées par dialogue; en attribuant à chacune des variables trois modalités dont les pivots étaient définis respectivement par les rangs suivants:

Tflx {2,8,15}; Tvll {2,8,15}; Rtrc {2,9,15}; Rflx {4,8,15}; Rchc {3,8,15}.

(par exemple, pour Rchc, pivot 2 = 84, valeur atteinte pour pes).

Ce codage nous paraissant peu fidèle au niveau de certaines valeurs extrêmes, on l'a modifié en reprenant sur éditeur de texte le fichier 'Dcodx' des valeurs pivot créé lors du premier passage. On est alors rentré dans 'zrang' en demandant de découper les variables avec des bornes (pivots) préétablies.

Ainsi, par exemple, avec les pivots {120,210,300} pour Tflx, on a:
pa6 = { 1, 0, 0}; pek = {0.66 , 0.33 , 0};
ce qui nous paraît donner à pa6 une excenticité raisonnable. Le tableau 'Dcodx' est publié à la fin de l'article.



3.3 Résultats de l'analyse factorielle

trace :	1.175e+0										
rang :	1	2	3	4	5	6	7	8	9	10	
lambda :	4249	1945	1673	1430	903	737	484	203	103	24	e-4
taux :	3616	1655	1424	1217	769	627	412	173	88	20	e-4
cumul :	3616	5271	6695	7912	8680	9307	9719	9892	9980	10000	e-4

On voit d'abord que toute l'inertie est distribuée sur 10 axes: ceci était attendu parce que le tableau a 15 colonnes (5 × 3), formant 5 blocs de 3 dont chacun a même total, ce qui réduit le rang à 10 (5 - 3).

En projection sur l'axe 1 on rencontre successivement les modalités 1 (F1<0) puis 2 (F1≈0) et enfin 3 (F1>0): ceci montre qu'on a un facteur de niveau général. Le facteur 1 peut servir de mesure globale de la qualité des produits. De fait, les produits non renforcés, dont les sigles sont en minuscules, sont généralement du côté (F1<0); tandis que les produits renforcés de fibres de verre, dont les sigles sont en majuscules, sont généralement du côté (F1>0). Il y a 3 exceptions: F1(PS) est négatif, mais proche de 0; pcl et pai sont de très bons produits et ont donc (F1>0); l'article souligne d'ailleurs que les polymères à cristaux liquides (pcl) ont des performances élevées même sans renfort.

Il vaut la peine d'interpréter quelques détails du plan (1,2). Avec une valeur maxima de F2, pcl apparaît associé à d'excellentes résistances (Rtr3, Rch3); mais sa tenue aux fortes températures n'est pas bonne: ce qui contribue à rendre positifs F2(Tvl1) et F2(Tfx1). Les matières PEEK, PCL et PPS, situées à l'extrémité positive de l'axe 1 sont, dans l'ensemble excellentes; mais leurs résistances au choc et à la traction sont moyennes; ce qui les éloigne de Rtr3 et de Rch3.

```
Tww:Usine Nouvelle, Technologie, Janvier 1989, superplastiques.
HK:plD:plTDcodx: bornes pour le decoupage des variables
le nombre des variables est  5
Tflx a  3 modalites dont les sigles et valeurs pivot sont
           Tfx1           Tfx2           Tfx3
  1.200000000e+2  2.100000000e+2  3.000000000e+2
Tvll a  3 modalites dont les sigles et valeurs pivot sont
           Tvl1           Tvl2           Tvl3
  1.250000000e+2  1.800000000e+2  2.350000000e+2
Rtrc a  3 modalites dont les sigles et valeurs pivot sont
           Rtr1           Rtr2           Rtr3
  8.000000000e+1  1.400000000e+2  2.000000000e+2
Rflx a  3 modalites dont les sigles et valeurs pivot sont
           Rfl1           Rfl2           Rfl3
  2.800000000e+1  7.800000000e+1  1.280000000e+2
Rchc a  3 modalites dont les sigles et valeurs pivot sont
           Rch1           Rch2           Rch3
  5.500000000e+1  8.500000000e+1  1.500000000e+2
```

4 Conclusion

L'exemple qui fait l'objet du §3 ne comporte pas assez de données pour permettre d'acquérir une vue d'ensemble des nouvelles matières plastiques à hautes performances. Mais c'est précisément le petit nombre de données qui nous a contraint à utiliser le codage barycentrique et a été ainsi l'occasion de créer une nouvelle option dans le programme 'zrang'. Divers essais non rapportés ici nous permettent de penser que le codage adopté a permis de donner aux résultats une forme équilibrée, peu sensible aux aléas des choix et de l'échantillonnage.

En général, l'intérêt du codage pour l'analyse des données n'est plus à démontrer: il nous paraît donc utile d'accéder par un programme unique à des formes de codage aussi diverses que possible (cf. §3).

Bibliographie

Bien que de nombreux travaux utilisent les divers codages considérés dans cet article, nous croyons devoir citer particulièrement:

S. Chaïeb: Variation de la concentration plasmatique d'une substance au cours d'une perfusion et après celle-ci; cas du dinitrate d'isosorbide; *CAD*, Vol IX, n° 1, 1984.

F. J. Gallego : Codage flou en analyse des correspondances; *CAD*, Vol VII, n° 4, 1982.